IN THE CLAIMS

Claims 1, 6, 10, 12, 13, 14, 24, 28 and 36 are amended herein. Claims 2, 7, 15, and 26 are cancelled. All pending claims are produced below.

1. (Currently Amended) A system for finding compounds in a text corpus, comprising:

a vocabulary comprising tokens extracted from a text corpus; and

a compound finder configured to iteratively identify ~~identifying~~ compounds having a plurality of lengths within the text corpus, each compound comprising a plurality of tokens, comprising:

an iterator configured to select *n*-grams having a same length that is less than a length of *n*-grams selected during a previous iteration;

an *n*-gram counter configured to evaluate ~~evaluating~~ a frequency of occurrence for one or more *n*-grams having the same length in the text corpus, each *n*-gram comprising at least one token ~~tokens~~ selected from the vocabulary; and

a likelihood evaluator configured to determine ~~determining~~ a likelihood of collocation for one or more of the *n*-grams having ~~a~~ the same length, adding ~~the~~ a subset of *n*-grams having a high ~~highest~~ likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds.

2. (Cancelled)

3. (Currently Amended) A system according to Claim 1, wherein only some of the subset of *n*-grams having a high ~~highest~~ likelihood are added as compounds to the vocabulary.

4. (Original) A system according to Claim 1, wherein the likelihood of collocation as a likelihood ratio λ is computed in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

5. (Original) A system according to Claim 4, wherein the $L(H_c)$ is determined, comprising dividing the *n*-gram into *n*-1 pairings of segments, calculating a likelihood of collocation for each pairing of segments, and selecting the maximum likelihood of collocation of the pairings as $L(H_c)$.

6. (Currently Amended) A method for finding compounds in a text corpus, comprising:

building a vocabulary comprising tokens extracted from a text corpus; and

iterativel identifying compounds having a plurality of lengths within the text corpus, each compound comprising a plurality of tokens, comprising:

selecting *n*-grams having a same length that is less than a length of *n*-grams selected during a previous iteration;

evaluating a frequency of occurrence for one or more *n*-grams having the same length in the text corpus, each *n*-gram comprising at least one token ~~tokens~~ selected from the vocabulary;

determining a likelihood of collocation for one or more of the *n*-grams having ~~a~~ the same length; and adding ~~the~~ a subset of *n*-grams having a high ~~highest~~ likelihood as compounds to the vocabulary and rebuilding the vocabulary based on the added compounds.

7. (Cancelled)

8. (Currently Amended) A method according to Claim 6, further comprising:

adding only some of the subset of the *n*-grams having a high ~~highest~~ likelihood as compounds to the vocabulary.

9. (Original) A method according to Claim 6, further comprising~~:~~ computing the likelihood of collocation as a likelihood ratio λ in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

10. (Currently Amended) A method according to Claim 9, further comprising~~:~~ determining $L(H_c)$, comprising:

dividing the *n*-gram into *n*-1 pairings of segments;

calculating a likelihood of collocation for each pairing of segments; and

selecting the maximum likelihood of collocation of the pairings as $L(H_c)$.

11. (Original) A computer-readable storage medium holding code for performing the method according to Claim 6.

12. (Currently Amended) An apparatus for finding compounds in a text corpus, comprising:

means for building a vocabulary comprising tokens extracted from a text corpus; and

means for iteratively identifying compounds having a plurality of lengths within the text corpus, each compound comprising a plurality of tokens, comprising:

means for selecting *n*-grams having a same length that is less than a length of *n*-grams selected during a previous iteration;

means for evaluating a frequency of occurrence for one or more *n*-grams having the same length in the text corpus, each *n*-gram comprising at least one token ~~tokens~~ selected from the vocabulary;

means for determining a likelihood of collocation for one or more of the *n*-grams having ~~a~~ the same length; and

means for adding a subset of ~~the~~ *n*-grams having a high ~~highest~~ likelihood as compounds to the vocabulary and means for rebuilding the vocabulary based on the added compounds.

13. (Currently Amended) A system for identifying compounds through iterative analysis of measure of association, comprising:

~~a stored limit on a number of tokens per compound~~

an iterator initially specifying a limit on a number of tokens per compound for an iteration and decreasing the limit for a subsequent iteration; and

a compound finder configured to iteratively evaluate ~~evaluating~~ compounds within a text corpus, comprising:

an *n*-gram counter configured to determine ~~determining~~ a number of occurrences of one or more *n*-grams within the text corpus, each *n*-gram comprising ~~up to~~ a number of tokens up to the limit for the iteration ~~a maximum number of tokens~~, which are ~~each~~ at least in part provided in a vocabulary for the text corpus;

a likelihood evaluator configured to identify ~~identifying~~ at least one *n*-gram comprising a number of tokens equal to the limit for the iteration based on the number of occurrences and determining a measure of association between the tokens in the identified *n*-gram, ~~and~~ adding each identified *n*-gram with a sufficient measure of association to the vocabulary as a compound token~~,~~ and rebuilding the vocabulary based on the added compound tokens~~ and adjusting the limit~~.

14. (Currently Amended) A system according to Claim 13, further comprising:

a stored upper limit on a number of identified *n*-grams; and

a limiter identifying a number of *n*-grams up to the stored upper limit based on the number of occurrences.

15. (Cancelled)

16. (Original) A system according to Claim 13, wherein the measure of association between the tokens in the identified *n*-gram comprises a likelihood ratio λ.

17. (Original) A system according to Claim 16, wherein the likelihood ratio λ is calculated in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

18. (Original) A system according to Claim 17, wherein, for each pair of tokens, t_1, t_2, in the identified n-gram, the independence hypothesis comprises $P(t_2 \mid t_1) = P(t_2 \mid \overline{t_1})$ and the collocation hypothesis comprises $P(t_2 \mid t_1) > P(t_2 \mid \overline{t_1})$.

19. (Original) A system according to Claim 17, wherein the $L(H_i)$ is computed for each pair of tokens, t_1, t_2, in the identified n-gram in accordance with the formula:

$$\arg\max_{L(H_i)} \frac{L(t_1, t_2 \mathit{form\ compound})}{L(n - \mathit{gram\ does\ not\ form\ compound})}.$$

20. (Original) A system according to Claim 13, further comprising:
an initial vocabulary comprising a plurality of tokens extracted from the text corpus.

21. (Original) A system according to Claim 20, further comprising:
a parser parsing the tokens from the text corpus.

22. (Original) A system according to Claim 13, further comprising:
a filter determining the number of occurrences of one or more n-grams within the text corpus for only unique n-grams.

23. (Original) A system according to Claim 13, wherein each text corpus comprises a plurality of documents comprising one of a Web page, a news message and text.

24. (Currently Amended) A method for identifying compounds through iterative analysis of measure of association, comprising:

iteratively specifying a limit on a number of tokens per compound for an iteration and decreasing the limit for a subsequent iteration;

~~specifying a limit on a number of tokens per compound~~; and

iteratively evaluating compounds within a text corpus, comprising:

determining a number of occurrences of one or more *n*-grams within the text corpus, each *n*-gram comprising up to a number of tokens up to the limit for the iteration ~~maximum number of tokens~~, which are ~~each~~ at least in part provided in a vocabulary for the text corpus;

identifying at least one *n*-gram comprising a number of tokens equal to the limit for the iteration based on the number of occurrences and determining a measure of association between the tokens in the identified *n*-gram; ~~and~~

adding each identified *n*-gram with a sufficient measure of association to the vocabulary as a compound token~~,~~ and rebuilding the vocabulary based on the added compound tokens~~ and adjusting the limit~~.

25. (Original) A method according to Claim 24, further comprising:

providing an upper limit on a number of identified *n*-grams; and

identifying a number of *n*-grams up to the upper limit based on the number of occurrences.

26. (Cancelled)

27. (Original) A method according to Claim 24, wherein the measure of association between the tokens in the identified n-gram comprises a likelihood ratio λ.

28. (Currently Amended) A method according to Claim 27, further comprising~~:~~ calculating the likelihood ratio λ in accordance with the formula:

$$\lambda = \frac{L(H_i)}{L(H_c)}$$

where $L(H_i)$ is a likelihood of observing H_i under an independence hypothesis, $L(H_c)$ is a likelihood of observing H_c under a collocation hypothesis, and H is a pair of tokens.

29. (Original) A method according to Claim 28, wherein, for each pair of tokens, t_1, t_2, in the identified n-gram, the independence hypothesis comprises $P(t_2 \mid t_1) = P(t_2 \mid \overline{t_1})$ and the collocation hypothesis comprises $P(t_2 \mid t_1) > P(t_2 \mid \overline{t_1})$.

30. (Original) A method according to Claim 28, further comprising:
computing the $L(H_i)$ for each pair of tokens, t_1, t_2, in the identified n-gram in accordance with the formula:

$$\arg\max_{L(H_i)} \frac{L(t_1, t_2\, form\ compound)}{L(n - gram\ does\ not\ form\ compound)}.$$

31. (Original) A method according to Claim 24, further comprising:
constructing an initial vocabulary comprising a plurality of tokens extracted from the text corpus.

32. (Original) A method according to Claim 31, further comprising:
parsing the tokens from the text corpus.

33. (Original) A method according to Claim 24, further comprising:

determining the number of occurrences of one or more *n*-grams within the text corpus for only unique *n*-grams.

34. (Original) A method according to Claim 24, wherein each text corpus comprises a plurality of documents comprising one of a Web page, a news message and text.

35. (Original) A computer-readable storage medium holding code for performing the method according to Claim 24.

36. (Currently Amended) An apparatus for identifying compounds through iterative analysis of measure of association, comprising:

means for specifying a limit on a number of tokens per compound for an iteration and decreasing the limit for a subsequent iteration ~~specifying a limit on a number of tokens per compound~~; and

means for iteratively evaluating compounds within a text corpus, comprising:

means for determining a number of occurrences of one or more *n*-grams within the text corpus, each *n*-gram comprising up to a number of tokens up to the limit for the iteration ~~maximum number of tokens~~, which are ~~each~~ at least in part provided in a vocabulary for the text corpus;

means for identifying at least one *n*-gram comprising a number of tokens equal to the limit for the iteration based on the number of occurrences and means for determining a measure of association between the tokens in the identified *n*-gram; and

means for adding each identified *n*-gram with a sufficient measure of association to the vocabulary as a compound token~~,~~ and means for rebuilding the vocabulary based on the added compound tokens ~~and means for adjusting the limit~~.